Hello World

一、机器学习简介

> 三大类别：监督学习、无监督学习、增强学习。

1.1 监督学习

目的是通过标注好的数据进行模型训练，从而期望利用训练好的模型对新的数据进行预测或者分类。

"监督"意味着已经有标注好的数据。

常见场景：垃圾邮件过滤、房价预测、图片分类等。弱点就是需要大量标注数据，前期投入大。

1.2 无监督学习

无监督相比监督学习，无需标注就可以达到某个目的。

常见场景：

- 聚类。如图片分类，可以预设一个类别总数进行自动划分即半监督学习；也可以预设一个差异阈值，对所有图片进行自动聚类。
- 降维。在数据特征多、维度过高的时候，需要把高维降到合理的低维空间处理，期望保留最重要的特征数据。主成分分析（PCA）就是常见的。

1.3 增强学习

监督学习和无监督学习，基础都来自于数据本身。增强学习最大的特点就是需要与环境有互动，也促使人们在增强学习的研究中利用类似电子游戏的环境来模拟互动进行AI训练。

二、核心开发流程

核心开发分为四段。

2.1 数据预处理

第一阶段就是处理原始数据。处理带有标签的原始数据，形成用于模型训练的训练数据集和验证模型的测试数据集。

2.1.1 特征提取

从原始数据挑选出来进行转换，并最终用于机器学习的数值就称为特征值。

2.1.2 数据清洗

目的是让算法用到的数据集尽量理想化，不包含不必要的干扰数值，从而提高模型训练的精度。

2.2 学习

也即是训练阶段。需要根据自己的最终目标选择合适的算法模型，并根据我们的数据集进行合理的参数设置，开始模型训练。

如多元函数：

y是标签，在数据集中已经标注好, 就是特征值。

2.3 评价

学习阶段将误差降到足够小之后，就可以停止训练，将训练好的模型用在数据预处理阶段生成的测试数据集上验证效果。

测试数据集中的所有数据都没有在训练阶段出现过，所以我们可以把测试数据作为新数据，模拟真实环境的输入，从而预估模型被部署到真实环境的效果。

2.4 预测

评估阶段确认模型达到了预期的准确率和覆盖率之后，可以部署上线。小规模研究中可以直接使用训练后的数据，但在真实流量的产品环境中，往往需要专门的模型服务框架，将模型转换为专有的格式，并在该框架进行高效服务。